Research Synopsis

DATA MINING TECHNIQUES: DEVELOPMENT FOR PROGNOSTICATION OF ACADEMIC INTERVENTION DPAI FOR HIGHER EDUCATION

Submitted By

Mr. Pawan Siddhartha Wasnik

School of Computational Sciences,

S. R. T. M. University, Nanded.

Under the Guidance of

Dr. S. D. Khamitkar

Associate Professor,

School of Computational Sciences,

S. R. T. M. University, Nanded.

Submitted to

Swami Ramanand Teerth Marathwada University, Nanded.

Data Mining Techniques: Development for Prognostication of Academic Intervention DPAI for Higher Education

Abstract. The emerging field of academic analytics that is educational data mining is rapidly producing new possibilities for gathering, analyzing and presenting educational data by extending their capabilities to discover hidden patterns in databases. In these lines, the objective of this research work is to introduce data mining over student's collected data as well as internationally known standard datasets. The ultimate goal is to find attributes in terms of set of invisible variables which severely affect the student's performance. Here data mining is intentionally used as the most common causes affecting performance within the datasets which cannot be seen normally. The results enable us to use discovered characteristic or patterns in palpating student's learning outcomes. These patterns can be useful for teachers to identify effective prognostication strategies for students.

1. Introduction

Student's data is a vital resource for all higher educational organizations. In today's era, Computer / Electronic / IT / ICT devices are used for data collections which help for collected student's data to be available in digital format. Due to such digitization, data can be processed easily leading to faster decision processes [16]. The role of data driven intelligence is becoming central in educational organizations. Contrary Data mining or Knowledge Discovery in Database (KDD) is the process of finding hidden and useful knowledge from large amount of data [1, 16]. It is an automated the process leading to accurate decision making. Now days, all educational organizations, institutions or Universities have been computerized and they have database systems capturing all essential data from all vital parts. Even though, higher educational institutes face very hard for predicting accurate knowledge from databases. Thus Data mining finds applications in educational industry [16] mainly to explore hidden knowledge in crucial aspects including student support, course registration processes, alumina associations, designing new courses, etc. We can apply Data Mining to traditional processes also. Such applications are very challenging [17]. These challenges can be addressed by use of Educational data mining algorithms. In Educational data mining, we primarily use existing data mining algorithms over educational databases to give good insights [2, 3, 16]. These insights are in terms of patterns. The discovered patterns can be used for characterizations of models for prediction, association, grouping, classification and better resource management [17]. These models are also known as Academic analytics. Despite the Government, UGC, etc are pouring lot of funds for improving education standards, academic performance of students from rural Universities of India is not improving. [1,3]. Over the years, University like Swami Ramanand Teerth Marathwada University, Nanded have taken attempts to create large amount of databases from the data pertaining to academic activities, educational

activities and administration contexts. If we apply educational data mining algorithms then it will be very interesting to see hidden patterns in these databases. Our work is the first most attempts. This work is an example of a joint interdisciplinary work undertaken by three Schools of our University, School of Computational Sciences, School of Mathematical Sciences and School of Educational Sciences. The objective of this study is to introduce Educational data mining processes to collected datasets. Using these results, one can realize what data has to be captured, what surrounding information has also to be captured, how mining can be carried out, etc for accurate insight. Thus the research objective of proposed work is to propose a strategy for automatic gaining of intelligence from vast educational data. In this research we will suggest approaches for finding patterns in student's achievements and performances. Some real word data related to performance of University students was collected using questionnaire and progress reports. A student's dataset was created with thousands of records and number of fields by closed questionnaire method. This includes student's personal details like social, intellectual, demographic, habitual, health and economical data.

2. Research Methodology

In India, the educational system consists of 10+2+ 3 pattern of undergraduate education for no-professional streams and 10+2+4 pattern for undergraduate professional streams. The students admitted to Computational Sciences allied subjects are mainly from the first pattern. Most of the students are under graduate from the public, grant-in-aids sanctioned and nominal expensive education system. There are several courses like BCS, BCA, B.Sc (CS or IT or CM) under science and technology faculty from where the students mainly come. Their data will be collected for the past few years. The database was built from two sources: previous examination's progress reports and our questionnaires.

We will show that the performance degrading pattern we have discovered will be helpful to predict some students which will also have failure. Our questionnaire will have questions with predefined options [18] related to student's attributes as defined in Pritchard and Wilson [5, 18]. The attributes in questionnaire have relations with the performance of students. The questionnaire will be reviewed by experts from Educational Sciences. There will be some trial testing of questionnaire on some small groups of students. These trials will help us to understand over all feedback as such. Some revisions with trial testing will be made to devise out the final version of questionnaire. The final questionnaire consisted of finite number of closed type questions [18]. The questionnaire will be distributed to students and demonstrated for feedback. After this, a dataset will be integrated with thousands of student records and each record consists of many fields. Originally, there will be fields equal to the total number of questions in questionnaire. Some additional fields will be added for seeking information of students. Microsoft Excel 2007 software can be used to record the dataset. Data set

values like Yes / No were converted in to numeric values like 1 0r 0. Other numerical codes in the range 0 ,1,2,3,4 …6 will also given depending upon the number of possible answers a question can have. Likewise other answers are also converted into numeric values. A sample questionnaire is given in figure 1. During the pre-processing of data, we may find some ambiguities and false information in student's data. We forecast around 10% confusing data or falsely filled data or partly filled questionnaire in first step. In order to correct them, the students will be called on one to one basis and convinced to fill the missing data. There can be some confidential issues discovered after going through the filled questionnaire. Appropriate actions will be taken time to time and significant number of students will get convinced.

Since we aim for discovery of attributes which affects performance of students, we primarily investigated literature across the globe to see what other people have done. To understand performance of a student, we underwent discussions with educationalist. The performance in broader sense is how well a student does in over all courses. For proper understanding the performance terminology, we primarily relied on the work of Shoukat Ali at al.[6]. This work is a lucid discussion on performance analysis of students. This work is base for our analysis and some terminology is also borrowed from the same. As per above refereed studies, there are numerous factors which affect student's academic growth and learning outcome. This consideration matches with few attributes of our dataset. In the literature review Shoukat Ali at al. [6] cites other references for proper understanding of factors important for performance analysis. A similar works of Graetz et al. [7] suggested that the social status of parents and performance of students are reliant with each other. This also matches with our understanding of proportionate relationship of performance with social and economical conditions of students. The Shoukat Ali et al.[6] also cites that Considine and Zappala [8] have noticed that parent's economical conditions have positive effects on the performance. The work of Staffolani and Bratti [9] highlights that the future achievements of students have correlation with previous year's scores.

1	Course code	MSc (5) , MCA (6)
2	Your name	
3	Gender (sex)	Male (1) Female (0)
4	Marital status	Married (2) unmarried(3)
5	Age	
6	Home address	Urban(1) rural (2) foreign(3)
7	Mobile no.	
8	Personal email id	
9	Degree passer and percentage	General B.Sc. / (1) B.Sc.(computer CS) / (2) BCA / BCS/ (3) Other / (4) (5)
	Percentage	
10	Degree collage name	
11	Father's Education	Below or SSC/ (1) HSC/ (2) Graduate/ (3) Post Graduate/ (4) other (5)
12	Fathers job and annual income	Service / (1) Business/ (2) Agriculture/ (3) In house/ (4) Other/ (5)
	Income	0-1 lakh (1) , 1.1-2 lakh(2), 2.1-5 lakh(3) , 5lakh - above (4)
13	Mothers education	Below or SSC/ (1) HSC/ (2) Graduate/ (3) Post Graduate/ (4) other (5)
14	Mothers job and annual income	Service / (1) Business/ (2) Agriculture/ (3) In house/ (4) Other/ (5)
	Income	0-1 lakh (1) , 1.1-2 lakh(2), 2.1-5 lakh(3) , 5lakh - above (4)
15	Family size	
16	Family relationship	Excellent / (1) Good/ (2) Satisfactory/ (3) Bad/ (4) Very Bad (5)
17	Family support to your education	Excellent / (1) Good/ (2) Satisfactory/ (3) Bad/ (4) Very Bad (5)
18	Reason to choose this course	Career in IT/ (1) Near to Home/ (2) Reputation of course / (3) Blind Decision/ (4) Parents wish: (5)
19	Travel mode and time needed	Bus/ (1) Railway/ (2) City Bus/ (3 but taken as 1) Rickshaw/ (4) Self Vehicle / (5) walking (6)

Fig.1. Sample Questionnaire

The combined finding that affects performance as per [6] highlights many social, economical and self related factors that have effect on student's academic performance. We felt that, once we discover these

hidden truths, we can undertake appropriate, corrective actions to improve overall performance. This is the motto behind our prognostication work. Further, it was matter of curiosity that whether such investigations were made in past or not. So we started investigating profiling studies for student's performances. In effect, several studies have been found which have addressed related issues. Some of them includes, Ma et al. [9] which used Association Rules [18], the Minaei-Bidgoli et al. [10] used Classification technique, the Kotsiantis et al. [11,18] used Prediction approach, and more recently, Pardos et al. [12,18] used Regression technique for prognostication purpose. Keeping the knowledge earned after careful review of above primary sources, in this work, we aim to predict attributes which are directly related to student's performance. We are of the opinion that not only productive teaching / learning but also other social and economical aspects will hamper the performance. Our analysis will carried out with some thoughts in mind including social aspects of students, economical favorableness of students to peruse education, personal desires & ambitions and finally resources available with students which can help them to get success.

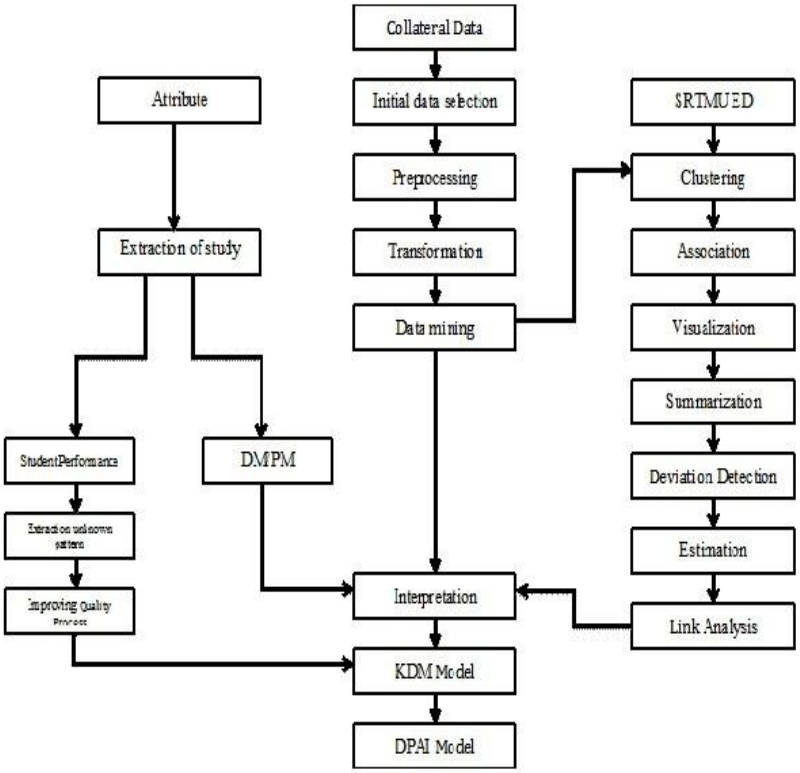

Fig. Research Methodology

3. Experimentations

We will use legal SPSS version 22 for the implementation of our idea[4]. Our student's dataset in MS-Excel can be exported in SPSS . There are other Open sources Platforms like R Miner, WEKA, etc for implementation of our experiments. We will use following data mining algorithms.

1. Association: Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as *relation technique*. The association technique is used in *market basket analysis* to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for customer and increase sales.
2. Classification: Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.
3. Clustering: Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.
4. Prediction: The prediction, as its name implied, is one of a data mining techniques that discover the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

5. Sequential Patterns: Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.
6. Decision trees: The A decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

   We can also combine two or more of those data mining techniques together to form an appropriate process that meets the business needs.

4. Expected Outcome

It was openly understood that many social, habitual and economical aspects are associated with performance of students and mere studying cannot be the sole criteria for good performance. However these are hidden and no attempt was made at to scientifically visualize them. The study took it as challenge. Using interdisciplinary approach, we will collect data and data mining algorithms will be implemented to get such insight. The specific objective of the undertaken research work is to find out any hidden patterns in the available data which could be useful for predicting students' performance. We will scientifically made visible the hidden aspects which affects the performance. The ultimate goal of our study is in terms of prognostication actions to improve academic performance of our students in their weaker aspects.

References

1. Margaret Dunham, Data Mining: Introductory and Advanced Topics, by Margaret H. Dunham, Pearson publications, 2002.

2. Han,J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray,2006.

3. Behrouz.et.al., Predicting Student Performance: An Application of Data Mining Methods with the Educational Web-Based System Lon-CAPA, IEEE, Boulder,

4. IBM SPSS Statistics 22 Documentation on internet retrieved at www.ibm.com/support/docview.wss?uid=swg27038407.

5. Pritchard, M. E., and Wilson, G. S. (2003). Using emotional and social factors to predict student success. Journal of College Student Development 44(1): 18–28.
6. Shoukat Ali et al , Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus, American Journal of Educational Research, 2013 1 (8), pp 283-289.

7. Graetz, B. (1995), Socio-economic status in education research and policy in John Ainley et al., Socio-economic Status and School Education DEET/ACER Canberra., J Pediatr Psychol. 1995 Apr;20(2):205-16.

8. Considine, G. & Zappala, G. (2002). Influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38, 129-148.

9. Bratti, M. and Staffolani, S. 2002, 'Student Time Allocation and Educational Production Functions', University of Ancona Department of Economics Working Paper No. 170.

10. Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000), Targeting the right
Students using data mining, Sixth ACM SIGKDD International Conference, Boston, MA (Conference Proceedings; p. 457-464.

11. B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and, W. F. Punch, Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA, in Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE, 2003.

12. Kotsiantis S. 2009. Educational Data Mining: A Case Study for Predicting Dropout – Prone Students. Int. J.Knowledge Engineering and Soft Data Paradigms, 1(2), 101–111.

13. Pavel Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, available at www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf.

14. K.Sasirekha, P.Baby, Agglomerative Hierarchical Clustering Algorithm- A Review , International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013 1 ISSN 2250-3153.
15. Nikhil Rajadhayx et al , Data mining in Educational Domain , retrieved from http://arxiv.org/pdf/1207.1535.pdf.

16. Gordon Linoff, Michael J, et al , Data Mining Techniques , 3e, Wiley Publications.

17. Eko Indrato , edited notes on Data Mining, retrieved from www. Http://recommender-systems.readthedocs.org/en/latest/datamining.html.

18. Paulo Cortez and Alice Silva ,Using Data Mining To Predict Secondary School Student Performance , retrieved from http://www.researchgate.net/ publication/ Using_data_mining_to_ predict_s econdary_ school_ student_ performance.

19. Keerthiram Murugesan, Jun Zhang , Hybrid Hierarchical Clustering: An Experimental Analysis , Technical Report: CMIDA-hipsccs #001-11, retrieved from www.cs.uky.edu/~jzhang/pub/techrep.html.

Research Guide
Dr. S.D.Khamitkar
Associate Professor
School of Computational Sciences.

Research Scholar
Mr. Pawan S. Wasnik
School of Computational Sciences.